MODEL-AGNOSTIC PREDICTABILITY QUANTIFICATION OF MULTIVARIABLE TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multivariable time series forecasting plays an important role in various domains, where predictions depend on complex interactions among covariates. Forecasting models are commonly evaluated across numerous benchmark datasets. However, forecasting accuracy reflects both model capability and task predictability, making it difficult to characterise the difficulty of a task independently of the model used. Existing predictability measures often fail to capture nonlinear dependencies or are limited to univariate series. In this work, we propose **M**ultiv**a**riable **P**redictability **S**core (MAPS), a model-agnostic metric for measuring the intrinsic predictability of multivariable time series forecasting tasks. MAPS uses covariate information to assess predictability across different lookback windows and forecasting horizons, and provides a quantitative assessment of the predictive information available for forecasting. We evaluate MAPS across synthetic and real-world forecasting datasets against 6 baseline predictability measures. The former confirms that our method closely aligns with the expected reference predictability ($r = 0.963$). Empirical results further corroborate that MAPS yields more reliable projections in the majority of experimental configurations.

1 INTRODUCTION

Time-series forecasting plays an important role across a wide range of domains and supports decision-making across multidisciplinary applications. Most forecasting problems are inherently multivariable. A single target variable is predicted from a set of covariates such as sensor information (Gneiting et al., 2006; Inman et al., 2013), economic indicators (Sims, 1980; Stock and Watson, 2002a;b), or clinical measurements (Lehman et al., 2014; Henderson et al., 2000; Wulfsohn and Tsiatis, 1997). However, forecasting tasks differ considerably in the predictive information they contain. Some exhibit strong temporal regularities and stable relationships between covariates and targets, whereas others are dominated by noise or weak predictive signals. This motivates a quantitative assessment of task predictability prior to modelling. Such an assessment can indicate whether a dataset supports reliable forecasting at a given horizon and provide insight into the performance that may be achievable.

Prior studies have attempted to characterise predictability in time series. An early approach (Sugihara et al., 1994) uses a forecasting method that predicts target values from similar historical patterns and relies on the forecasting accuracy to distinguish structured dynamics from stochastic noise. Zadeh et al. (2025) use mutual information to measure the relationship between inputs and targets and how much of this relationship remains in a model's prediction errors. Beck et al. (2025) assess predictability by testing whether forecasting models consistently outperform a naïve persistence model, considering significant reductions in forecast error as evidence of predictable structure in the data. However, these approaches are model-dependent, meaning that the resulting predictability estimate depends on both the data and the forecasting model rather than reflecting the intrinsic aleatoric property of the prediction task. Early approaches from information theory deploy entropy to assess predictability. Approximate entropy (Pincus, 1991) measures the extent to which recurring patterns in a time series remain similar, where a lower value indicates that the series is more predictable. Permutation entropy (Bandt and Pompe, 2002) uses Shannon entropy to measure the irregularity of changes in a time series, with lower values indicating higher predictability. However, neither method is formulated for multivariable forecasting tasks or considers how predictability changes across different forecasting horizons. Alternative approaches utilised the frequency domain. Powell and Percival (1979) introduced spectral entropy, which measures how consistently a time

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

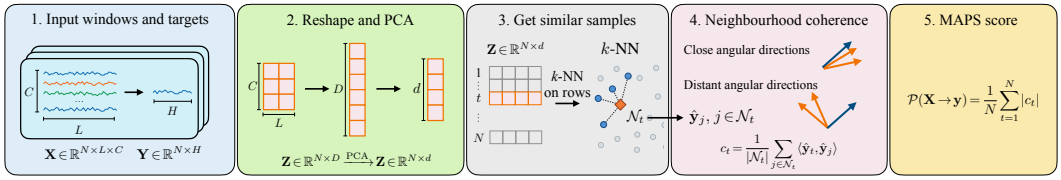


Figure 1: Overview of the proposed Multivariable Predictability Score (MAPS).

series follows repeating cycles, with lower values indicating clearer patterns and greater predictability. Building on this, ForeCA (Goerg, 2013) extends to multivariate settings by forming weighted sums of variables and identifying those with lower spectral entropy as more predictable. More recently, Wang et al. (2025a) transformed time series into the frequency domain and used distances between their frequency components as an indicator for forecasting model selection. Despite their computational efficiency, these frequency-domain approaches either consider a univariate setup (Powell and Percival, 1979; Wang et al., 2025a), or do not account for nonlinear dependencies (Goerg, 2013).

Some prior work accounts for the forecast horizon when assessing predictability. Wang et al. (2025b) measures how much frequency patterns vary across time-series segments, including both the observed history and the target to be predicted, with greater variation indicating lower predictability. Similarly, Catt (2026) assesses predictability at each forecast horizon using the mutual information between the observed past and the target value, with lower mutual information indicating less predictive information at that horizon. However, these approaches are designed for univariate time series and do not account for covariate dependencies. Furthermore, mutual information estimation in high-dimensional spaces is known to be computationally prohibitive (Kraskov et al., 2004), with complexity scaling poorly as the number of variables increases (Gao et al., 2015) (Figure S6).

Overall, existing predictability measures either depend on forecasting models, ignore forecasting horizons, or are limited to univariate settings. This gap is particularly relevant to the evaluation of forecasting models across benchmarks. Common long-term forecasting evaluations repeatedly use a small collection of datasets, including ETT, Electricity, Traffic, Weather, Exchange-Rate, and ILI (Zeng et al., 2023), while broader archives have mostly focused on collections of univariate series (Godahewa et al., 2021). Recent benchmarks such as ProbTS (Zhang et al., 2024), GIFT-Eval (Aksu et al., 2024), and fev-bench (Shchur et al., 2025) have expanded evaluation across domains and multivariate settings. Nevertheless, benchmark tasks can still differ widely in the predictive information they contain. Some exhibit strong recurring patterns that can be captured from the target history, whereas others rely on cross-variable relationships or contain only weak predictive signals. A model-independent measure that considers the available covariates, lookback window, and forecast horizon can help interpret performance beyond conventional forecasting metrics.

Motivated by the aforementioned observations, our main contributions are:

1. We propose **M**ultiv**a**riable **P**redictability **S**core (MAPS), a model-agnostic measure that quantifies the predictability of a multivariable forecasting task by assessing whether similar historical samples lead to consistent future target vectors. MAPS explicitly accounts for the available covariates, lookback window, and forecasting horizon without requiring a forecasting model.
2. We evaluate MAPS on synthetic time series with diverse dynamics and known predictability. MAPS closely follows the estimated predictability, achieving an average linear correlation of $r = 0.963$.
3. We compare MAPS against 5 existing predictability measures on multivariable Weather and Influenza-Like Illness forecasting tasks across multiple lookback windows and forecasting horizons. MAPS shows the strongest relationship with empirical forecasting performance, obtaining correlations of $r = 0.961$ on Weather and $r = 0.827$ on ILI.[1]

2 Predictability of Multivariable Time Series – Task Definition

We focus on quantifying the predictability of datasets for multivariable forecasting in a model-agnostic fashion. Specifically, we consider the multivariable forecasting task over a data set $\mathbf{Q} \in \mathbb{R}^{V \times C}$ where

[1]We note that these correlations are measuring different relationships which stem from the different experimental setups for these two forecasting tasks (see Section 4.4).

V denotes the number of observations and C is number of covariates, including 1 target and $C-1$ exogenous variables. From $\mathbf{Q}$ we extract N samples with a fixed stride of M. This forms an input tensor $\mathbf{X} \in \mathbb{R}^{N \times L \times C}$, where L is the lookback window length. The corresponding future target values are captured in $\mathbf{Y} \in \mathbb{R}^{N \times H}$, where H is the forecasting horizon. Each sample thus consists of an input $\mathbf{X}_t \in \mathbb{R}^{L \times C}$ covering the past time steps $\{t-L+1, \ldots, t\}$, and the corresponding target vector $\mathbf{y}_t \in \mathbb{R}^H$ covering the target time steps $\{t+1, \ldots, t+H\}$. The task is to define a predictability function $\mathcal{P}(\mathbf{X} \to \mathbf{Y})$ that quantifies the intrinsic predictability of the future target values from the corresponding input windows, independently of any specific model or supervised learning procedure.

3 Multivariable Predictability Score (MAPS)

An overview of the MAPS method is shown in Figure 1. We first standardise each variable (or column) of $\mathbf{Q}$ over its full time series. We then partition the standardised time series of each variable into N overlapping windows as described in section 2. $\mathbf{X}_t$ and $\mathbf{y}_t$ are additionally standardised to maintain local patterns within each sample.

Dimensionality Reduction with Principal Component Analysis (PCA). Following standardisation, each input window $\mathbf{X}_t$ is reshaped into vector $\mathbf{z}_t \in \mathbb{R}^D$, where $D = L \times C$. This forms the input state matrix held in $\mathbf{Z} \in \mathbb{R}^{N \times D}$. We note that the subsequent predictability computation operates on this D-dimensional representation and does not explicitly consider temporal adjacency. For multivariable datasets with many covariates or long lookback windows, D can be rather large. To reduce the dimensionality of this representation, we optionally apply PCA (Jolliffe, 1986) to project $\mathbf{Z}$ onto $d = \min(d, N-1, D)$ leading principal components, where d is an adjustable target dimensionality. For simplicity, we use $\mathbf{Z} \in \mathbb{R}^{N \times d}$ and $\mathbf{z}_t \in \mathbb{R}^d$ to denote the matrix and state vectors after this optional projection, respectively, with $d = D$ when PCA is not applied.

Identify Similar States with Nearest Neighbours. We use k-nearest neighbours (k-NN) to identify similar samples among the N sliced windows $\mathbf{Z}$. For each input sample $\mathbf{z}_t$, we identify its k nearest neighbours $\mathcal{N}_t$ using the Euclidean distance, excluding $\mathbf{z}_t$ itself. The neighbourhood size k determines a local component for the proposed predictability score. It should be large enough to provide statistically stable variance estimates, while small enough to preserve the local nature of the analysis. Following established practices in nearest-neighbour (Stone, 1977; Devroye et al., 1994; Levada et al., 2024), we set $k = \lfloor \log_2(N) \rfloor$.

3.1 Quantifying Dataset Predictability

Our method is motivated by local regularity in the relationship between observed states and their future evolution (Stone, 1977). In deterministic dynamical systems, the current state determines the future trajectory (Lorenz, 1963), while state-space reconstruction provides a basis for studying how nearby observations evolve (Takens, 1981; Eckmann and Ruelle, 1985). Accordingly, when the input contains predictive information about the target, nearby samples should correspond to target vectors with similar directions. We thus assess how consistently neighbouring samples align in the vector direction as an empirical measure of the predictive structure captured by the input samples.

Given the standardised target vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{R}^H$, we first obtain the corresponding unit direction vectors as $\hat{\mathbf{y}}_t = \frac{\mathbf{y}_t}{\|\mathbf{y}_t\| + \epsilon}$, where $\epsilon = 10^{-8}$ is a small constant that prevents division by zero. The resulting target vectors thus have unit length. The neighbourhood coherence between the current window t and a neighbouring window j is then obtained by computing the cosine similarity $\langle \hat{\mathbf{y}}_t, \hat{\mathbf{y}}_j \rangle \in [-1, 1]$. Averaging over $\mathcal{N}_t$, we have:

$$c_t = \frac{1}{|\mathcal{N}_t|} \sum_{j \in \mathcal{N}_t} \langle \hat{\mathbf{y}}_t, \hat{\mathbf{y}}_j \rangle \,, \tag{1}$$

where $c_t \in [-1, 1]$ measures the average directional agreement between window t and the futures of windows in $\mathcal{N}_t$, whose observed histories are most similar.

The predictability of a time series at window t is then given by $s_t = |c_t|$, where, by construction, $s_t \in [0, 1]$. Importantly, we take the absolute value of c_t, after averaging the directional similarities. This is because both positive ($c_t \to 1$) and negative ($c_t \to -1$) directional agreement of the target vectors are equally indicative of predictability. In this case, large $|c_t|$ values indicate that the k-nearest neighbours of window t follow coherent future directions, reflecting a more predictable

mapping between similar historical samples and their corresponding target vectors. By contrast, neighbourhoods containing orthogonal future directions yield $c_t \approx 0$, indicating that windows with similar observed histories do not evolve consistently. In such cases, the input representation provides limited information about future dynamics.

The overall predictability score of a time series dataset w.r.t. the specific look-back window and forecasting horizon is thus given by:

$$\mathcal{P}(\mathbf{X} \to \mathbf{y}) = \frac{1}{N} \sum_{t=1}^{N} s_t \in [0,1] \ . \tag{2}$$

Intuitively, $\mathcal{P}(\mathbf{X} \to \mathbf{y}) \approx 1$ indicates high predictability, where samples that are close in the input space have futures that also evolve similarly, indicating a more deterministic input-to-target mapping. Contrarily, $\mathcal{P}(\mathbf{X} \to \mathbf{y}) \approx 0$ indicates low predictability, i.e. even nearly identical samples result in diverged directions, suggesting chaotic dynamics that are potentially less predictable.

Computation cost. Performing PCA cost $O(ND\min\{N,D\})$, an exhaustive Euclidean nearest-neighbour search over N windows costs $O(N^2 d)$, and conducting cosine similarity over the neighbouring target vectors costs $O(NkH)$. Given d and H are fixed and $k = \lfloor \log_2 N \rfloor$, this complexity simplifies to $O(ND\min\{N,D\} + N^2)$. We additionally provide runtime comparison in Appendix C.6.

Constant target variable. A special case emerges when the target variable is a constant. Standardisation is applied over the entire sequence, so a constant target will become equal to zero in every window ($\mathbf{y}_t = 0$) and leave the direction undefined. Given that a constant target can be forecast exactly by its own value, we set $s_t = 1$ regardless. A near-constant target is otherwise scored normally as it does not collapse to zero vector. We further discuss this in Appendix C.1.

4 Results

We evaluate MAPS on both synthetic and real-world datasets and compare it with 6 baseline methods. We further examine its behaviour under distribution shift and robustness to different configurations.

4.1 Multivariable Time Series Datasets

Synthetic time series. To evaluate our method under controlled conditions, we construct synthetic datasets with controllable predictability. We begin by combining signals with known deterministic and stochastic components, including sinusoidal, linear trends, exponential and logarithmic growth, and Gaussian noise. By varying the proportion and magnitude of each component, we control the level of predictability. In this way, we form 6 basic synthetic data sets. We further consider 6 complex dynamic systems to enable a more comprehensive evaluation: a linear Gaussian state-space model, a smooth Gaussian process, a fractional Gaussian noise process, a two-state hidden Markov model, a noisy Lorenz system, and a quasi-periodic process. These cover a range of different predictability decay behaviours w.r.t. the forecasting horizon, enabling a horizon-dependent evaluation of our predictability metric (Dong et al., 2025).

Real-world tasks. We form the benchmark multivariable time series datasets from the Weather (`WEA`) and Influenza-Like Illness (`ILI`) forecasting tasks introduced by Shu and Lampos (2026). For `WEA`, the forecasting target is variable 't850' at a target location (for 20 locations in total), and the time step is set to 6 hours. We consider both a task that uses 5 covariates at the target location, and also expand this to a task with 45 covariates across a 3×3 grid of locations (target and 8 neighbouring ones). `ILI` contains daily ILI rates and corresponding Google search frequencies for related queries for USA HHS Regions 2, 9, and 10, and England, UK.

Further details for both synthetic and real-world datasets are provided in Appendix A.

4.2 Experiment configurations

Baseline methods. We compare MAPS against 5 model-agnostic baseline predictability methods. These are a persistence baseline, ForeCA (Goerg, 2013), Spectral Entropy (Wang et al., 2025a), Pattern Complexity (PC) (Wang et al., 2025b), and Mutual Information (MI). To adapt baselines

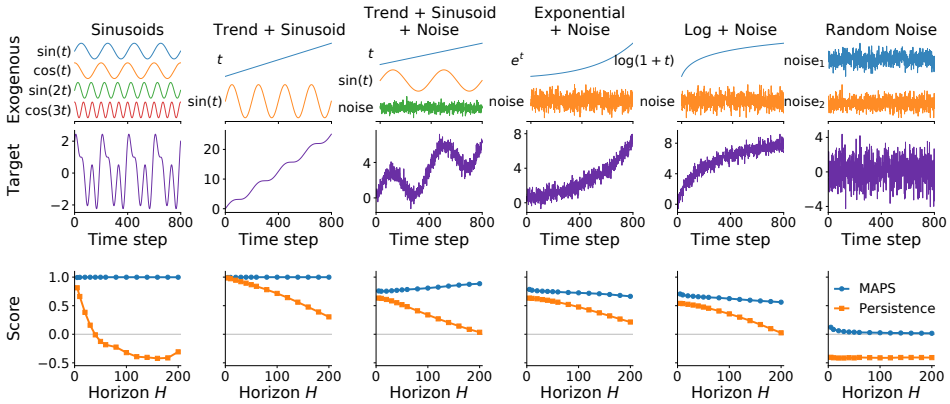


Figure 2: Basic synthetic time series used to validate MAPS. The exogenous variables are shown in the top row, the corresponding target series in the middle row, and the predictability w.r.t. H in the bottom row, each is a combination of the exogenous inputs. Predictability scores are computed using a rolling window where the input includes all covariates, i.e. both exogenous variables and the target.

that are originally designed for univariate forecasting to multivariable settings, we compute the scores for each input variable w.r.t. the target and average across variables. We additionally include Epiplexity (Finzi et al., 2026), a model-dependent measurement, where we use ridge regression as the predictive model in our experiments. Further details are provided in Appendix B.

Experiment setup. We conduct experiments with varying lookback window lengths (L) and forecasting horizons (H). Specifically, we consider $L \in \{28, 56, 84\}$, $H \in \{4, 12, 28\}$ for `WEA`, and $L \in \{28, 56\}$, $H \in \{7, 14, 21, 28\}$ for `ILI`. We obtain $d = 50$ principal components and set the neighbourhood size to $k = \lfloor \log_2 N \rfloor$. Unless otherwise stated, we sample rolling windows using a stride of $M = 10$.

Evaluation metrics. We use the linear correlation coefficient r between estimated and reference predictability scores as the primary metric in our evaluation. For synthetic data, the reference value is the designed predictability of each data-generating process. For real-world data, we use the performance of a competitive machine learning model as the reference. Machine learning performance is mainly determined by the linear correlation coefficient ρ between the model predictions and observed targets, a metric that is invariant to both scaling and translation. We additionally consider a scaled absolute forecasting error as a complementary measure. Specifically, let e denote the MAE of the model on the test set, and σ_y be the standard deviation of the target series over the test period. We first compute the standardised MAE e/σ_y (Gasteiger et al., 2020), which expresses absolute forecasting error in units of target standard deviation. We further transform this into the scaled MAE, $s = \sigma_y/(\sigma_y + e) \in (0, 1]$, with higher values indicating better forecasting performance.

Forecasting configuration. For the controlled synthetic systems, we use ridge regression as an interpretable forecasting baseline, with the penalty coefficient set to 1. For the real-world experiments, we use the performance of multivariable forecasting models as an empirical reference to evaluate the predictability scores. Predictability is computed on the training set, and forecasting performance is evaluated on the test set. Training settings and the Weather and ILI data splits are detailed in Appendix B.

4.3 Predictability under Controlled Conditions with Synthetic Data Sets

We commence by conducting experiments using 6 basic synthetic data sets over 800 time steps with known predictability and consider a forecasting setting of $L = 50$. The first two are deterministic, and their expected predictability score is 1. The $3-5$ series adds i.i.d. Gaussian noise to 3 deterministic functions, making the series less predictable. The sixth is only Gaussian noise and is therefore not predictable. Outcomes are depicted in Figure 2. MAPS assigns a score of 1 to both deterministic series across different H, about 0.77 to the series with added noise, and close to 0 to pure Gaussian noise, corroborating the expected predictability. Contrarily, persistence reflects the pattern of the series rather than its predictability.

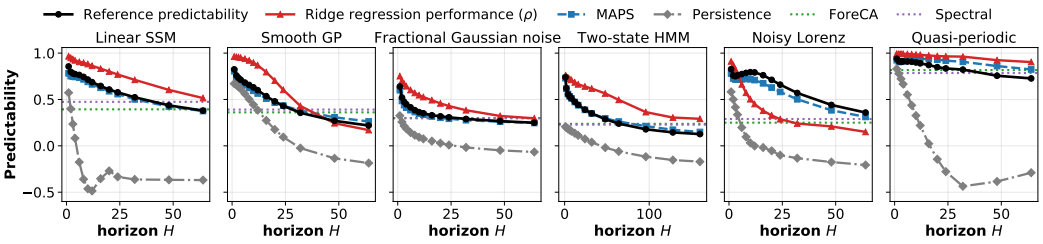


Figure 3: Reference predictability compared to MAPS, ForeCA, Spectral, Persistence, and the empirical performance (using linear correlation) of a ridge regression model for complex synthetic time series tasks across increasing forecasting horizons (H) for a fixed lookback window ($L = 16$).

We further evaluate MAPS on 6 synthetic univariate tasks that exhibit more complex and nonlinear characteristics and hence have different degrees of predictability. This univariate setting ($C = 1$) serves as a special case of the forecasting task considered by MAPS. For each task, we generate sequences of 4,000 time steps and compare a reference (expected) predictability score with MAPS w.r.t. different forecasting horizons (H), under a fixed look-back window of $L = 16$. We define the reference predictability as

$$P(L,H) = 1 - \frac{\mathbb{E}\left[\sum_{h=1}^{H} \mathrm{Var}(y_{t+h} \mid y_{t-L+1:t})\right]}{\sum_{h=1}^{H} \mathrm{Var}(y_{t+h})}, \tag{3}$$

where y_t is the generated future target value at time t. We use the known data-generating process to obtain possible future sequences given the observations in the lookback window and calculate how much uncertainty remains across these futures. $P(L,H) \in [0,1]$ quantifies uncertainty after observing past values within the look-back window L. A value of 1 indicates that the past fully determines the target value, while 0 dictates that the past provides no predictive information (see also Appendix A.2.1).

Figure 3 compares the reference predictability with MAPS, 2 frequency-based baselines (ForeCA and Spectral), persistence baseline, alongside the average correlation between predictions and ground truth for a ridge regression model as a baseline. It can be observed that MAPS strongly correlates with $P(L,H)$ ($r = .963$). Their scores actually overlap, with the exception of the quasi-periodic task where MAPS slightly overestimates predictability. Ridge regression performance estimates exhibit slightly weaker agreement with the target predictability than MAPS, but still a significant one ($r = .924$). This alignment further motivates the use of forecasting model performance as the reference predictability in the follow-up experiments. Persistence achieves a lower correlation ($r = .852$) and oftentimes underestimates predictability, particularly at longer forecasting horizons. Given this mediocre performance and the fact that its score reflects the error of the persistence predictor rather than data predictability, we do not report Persistence in the following experiments. Additionally, the two spectral baselines perform poorly on datasets without strong seasonality and fail to capture how predictability varies across different L/H settings.

4.4 Evaluating Predictability Against Empirical Forecasting Performance

We further evaluate predictability measurements on real-world data sets (`WEA` and `ILI`). For `WEA`, we first average predictability and performance scores across the test periods and then compute their correlation over 20 locations for each forecasting horizon. We perform this for two configurations, considering just the target location (5 covariates) or including a grid of 8 neighbouring locations (45 covariates). Results are shown in Figure 4 and Table 1. Across the compared methods, MAPS shows the strongest correlation with the performance (ρ) of Sonnet. Among the baselines, Epiplexity and ForeCA are the most competitive. However, Epiplexity is model-dependent, meaning that its effectiveness is closely related to the model used.

Additionally, including climate indicators from neighbouring regions (45 covariates) is expected to improve the reference performance ρ over using only local indicators (5 covariates) by providing additional predictive information. Importantly, MAPS reflects this improvement. The configuration

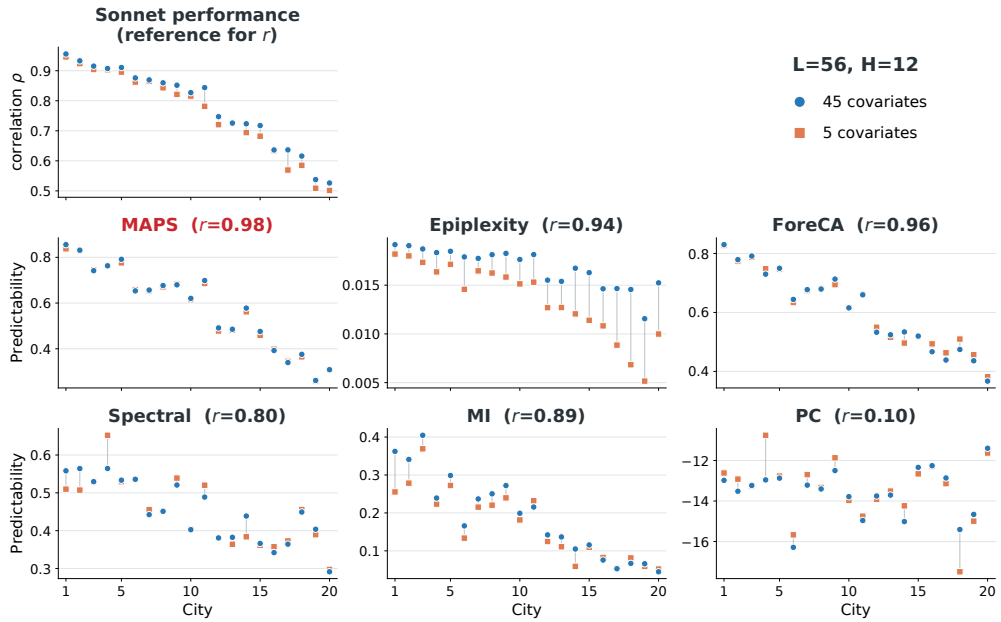


Figure 4: Predictability scores of MAPS and 5 baselines for the WEA task compared against the forecasting performance of Sonnet across 20 locations (lookback L=56, horizon H=12) for two configurations (5 or 45 covariates). Location order reflects Sonnet's correlation performance.

with more covariates yields higher MAPS scores. The baseline measures largely fail to capture this distinction, except for the model-based method Epiplexity and MI.

Results across all forecasting configurations of WEA are enumerated in Table 1. When we compare the predictability score with target correlation ρ, MAPS yields the strongest correlation r in 9 out of 18 configurations. It performs best in five of six configurations with $H = 12$ and in four out of six configurations with $H = 28$. Epiplexity is the most competitive baseline when referred to ρ, obtaining stronger results for the shortest forecasting horizon $(H = 4)$. Turning to the scaled-MAE score s, MAPS yields the strongest correlation in 13 out of 18 configurations. ForeCA is the most competitive baseline for s, achieving the highest correlation in six configurations, while Epiplexity leads in one. Notably, MAPS is more competitive at longer forecasting horizons. It leads or ties in five of six settings at $H = 12$ and four of six at $H = 28$ for target correlation ρ, and in five and six settings, respectively, for scaled MAE. MAPS's alignment with s is also significantly stronger than Epiplexity's (paired t-test, $p < 0.001$). The final two columns provide a complementary within-

Table 1: Average linear correlation (across all test periods) between predictability scores and Sonnet's forecasting performance, which is measured by correlation (ρ) or scaled MAE (s). Correlation breakdowns for lookback windows (L), forecasting horizons (H), and covariate number (C) are provided for the WEA task, over the 20 considered locations. The right-most two columns enumerate the average correlation among all combinations of (L, H) across regions for WEA and ILI.

	Method	WEA, $L=28$						WEA, $L=56$						WEA, $L=84$						Across (L,H) pairs	
		$H=4$		$H=12$		$H=28$		$H=4$		$H=12$		$H=28$		$H=4$		$H=12$		$H=28$		WEA	ILI
		$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$		
ρ	MAPS	.843	.896	**.975**	.971	.967	.972	.812	.872	**.977**	**.975**	**.972**	**.968**	.797	.856	**.975**	**.973**	**.973**	**.976**	**.961**	**.827**
	Epiplexity	**.991**	**.955**	.951	.951	**.968**	**.976**	**.987**	**.956**	.922	.953	.940	.967	**.961**	**.959**	.853	.949	.859	.974	.833	.455
	ForeCA	.801	.836	.972	**.974**	.961	.972	.804	.845	.976	.974	.970	.963	.805	.850	.974	.972	**.973**	.973	–	–
	Spectral	.459	.514	.844	.768	.851	.779	.471	.525	.836	.763	.840	.780	.467	.526	.836	.779	.848	.788	–	–
	MI	.741	.780	.902	.868	.906	.883	.718	.766	.910	.873	.914	.881	.711	.757	.905	.878	.928	.897	.470	.403
	PC	.350	.458	.036	.328	-.031	.309	.297	.436	-.057	.260	-.089	.237	.292	.452	-.120	.217	-.101	.237	.371	.335
s	MAPS	.921	**.946**	**.960**	.951	**.955**	**.952**	.913	**.939**	**.969**	**.959**	**.958**	**.961**	.905	.934	**.964**	**.958**	**.957**	**.960**	**.946**	**.817**
	Epiplexity	**.923**	.902	.907	.896	.944	.938	.909	.896	.887	.895	.917	.945	.873	.904	.822	.892	.846	.940	.833	.537
	ForeCA	.915	.936	.956	**.960**	.947	**.952**	**.922**	.938	.965	**.959**	.953	.955	**.921**	**.940**	.958	.956	.953	.954	–	–
	Spectral	.664	.666	.905	.813	.902	.834	.687	.668	.903	.819	.888	.838	.678	.672	.897	.822	.893	.832	–	–
	MI	.899	.888	.907	.861	.901	.871	.895	.885	.931	.878	.920	.891	.886	.874	.927	.876	.931	.901	.450	.481
	PC	.335	.427	.054	.305	-.038	.271	.215	.332	-.062	.208	-.097	.208	.168	.322	-.153	.161	-.130	.177	.357	.267

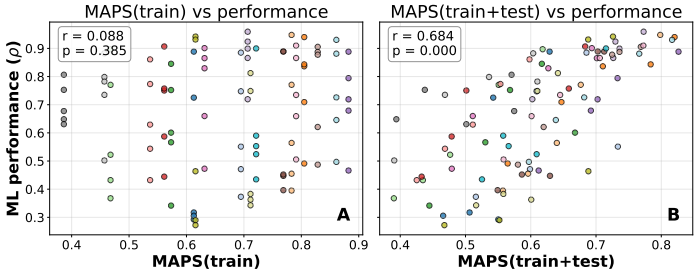


Figure 5: Correlation between predictability score and downstream forecasting performance across weather regions under distribution shift. (A) Predictability computed on the source (training) data. (B) Predictability computed on the combined training and test data.

Method	Train	Train+test
MAPS	0.088	0.684
Epiplexity	0.110	0.534
MI	0.067	0.523
Spectral	−0.099	0.531
ForeCA	0.108	0.519
PC	−0.109	0.381

Table 2: Linear correlation r between each predictability measure and the downstream forecasting performance across the same cross-location weather tasks.

location analysis across different L, H settings. Baselines that do not produce task-specific scores (ForeCA and Spectral) are marked "–" because their scores remain constant. MAPS achieves the highest within-location correlation in WEA/ILI, reaching $0.961/0.827$ with ρ, and $0.946/0.817$ with s. These results show that the alignment between MAPS and forecasting performance is consistent across datasets and (L, H) settings, capturing changes in forecasting difficulty across tasks.

4.5 Predictability Under Distribution Shift

We evaluate whether dataset predictability provides a useful prior for forecasting performance under real-world distribution shift. Specifically, we train a forecasting model on one region's weather data and evaluate it on another, constructing a diverse set of cross-location transfer tasks to approximate the distribution shift. To ensure a comprehensive evaluation, we train the forecasting model on each of the 20 locations and evaluate it on five other locations, selecting the two geographically nearest and three geographically farthest locations for each case.

Figure 5.A shows that the predictability score of the training data alone shows a weak relationship with downstream performance ($r = 0.088$). This is expected as the source dataset provides limited information about the predictability of the target series if there is a distribution shift. In contrast, when predictability is estimated on the combined training and deployment data (Figure 5.B), a slightly more positive correlation emerges ($r = 0.684$). Table 2 also shows that other baselines present a similar change. It can be concluded that the current predictability scores measure the intrinsic predictability of the time series on which they are computed, and they do so accurately once that distribution is observed. However, it should be noted that in deployment settings subject to shift, a score computed at training time would overstate or understate the difficulty of the task.

4.6 Ablation Study for MAPS

In this section, we perform an ablation study by replacing one component of MAPS at a time to justify our design choice. 'w/ variance' replaces the direction-based similarity measure with a

Table 3: Component ablation of MAPS. Correlations are averaged over all combinations of (L, H) across locations for WEA and ILI. Best results are **bolded**.

Method	WEA (r)	ILI (r)
MAPS	**0.961**	**0.827**
w/ variance	0.882	0.593
w/ signed directions	0.902	0.480
w/o standardisation	**0.961**	0.733
w/o window standardisation	**0.961**	0.811
w/o PCA	0.960	0.607

Table 4: MAPS predictability scores for England ILI under different PCA variance thresholds for each (L, H) task. "No PCA" uses all input variables, while the other rows retain the fewest principal components needed to explain 70%, 80%, or 90% of the variance.

	$L=28$				$L=56$			
PCA	$H=7$	$H=14$	$H=21$	$H=28$	$H=7$	$H=14$	$H=21$	$H=28$
No PCA	0.470	0.467	0.463	0.458	0.541	0.517	0.510	0.497
70%	0.578	0.568	0.561	0.556	0.607	0.598	0.593	0.589
80%	0.558	0.542	0.531	0.525	0.581	0.573	0.570	0.568
90%	0.552	0.564	0.567	0.572	0.619	0.638	0.648	0.653

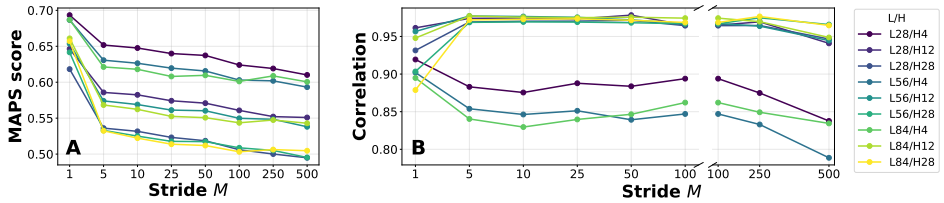


Figure 6: Effect of window stride on MAPS across nine (L, H) settings across 20 locations (WEA). (A) Averaged predictability score; (B) Correlation across 20 locations. The axis breaks at stride 100, which separates the equally spaced stride values from a linear scale over 100–500.

variance-ratio objective using the mean variance of the target and neighbouring features. Further detail in B.2. 'w/ signed directions' is instead of taking the absolute value, we map $[-1, 1]$ directly onto $[0, 1]$, so a negative alignment is mapped below 0.5. 'w/o standardisation' removes the z-score standardisation applied before window slicing. 'w/o window standardisation' removes the window-wise standardisation for each sliced window. Finally, 'w/o PCA' omits the PCA projection before the nearest-neighbour search. Table 3 shows how each design choice performs under controlled predictability scenarios and contributes to MAPS. MAPS score consistently achieves the strongest alignment with downstream forecasting performance on the real-world datasets, obtaining the highest correlation on both WEA ($r = 0.961$) and ILI ($r = 0.827$). Removing the standardisation has little effect, while removing PCA substantially degrades performance alignment on the ILI dataset (the dataset that contains more covariates), suggesting that dimensionality reduction improves the reliability of neighbourhood estimation in higher-dimensional input spaces. Replacing the neighbourhood direction alignment with a variance measurement or using signed directions also reduces agreement with forecasting performance.

4.7 Sensitivity analysis

Effect of PCA dimensionality. In MAPS, PCA is applied to remove redundancy among all available variables, thereby reducing the computational cost of the following k-nearest-neighbour search. This is particularly important for the ILI dataset, where each dataset contains 2k–5k covariates. Table 4 reports the MAPS score after retaining different proportions of the variance. The score is consistently higher after PCA than the no PCA version. We note that these results present as a sensitivity analysis of MAPS to the input. They do not, by themselves, show that a higher score is more accurate.

Effect of window stride. As shown in Figure 6.A, increasing the stride (M) lowers the absolute predictability score because adjacent windows overlap less, making neighbouring samples less similar. Despite this change in score magnitude, its correlation with downstream forecasting performance remains stable across a wide range of strides (Figure 6.B). Stride 1 performs slightly worse, as heavily overlapping windows may be selected as neighbours rather than revealing recurring patterns. Correlation begins to drop at larger strides (250 and 500), where too few windows remain to accurately capture the local structure of the dataset. Arguably, other factors such as dataset size and seasonality may also contribute to this decline, though we leave a closer analysis to future work. Overall, MAPS is robust to the choice of stride.

5 Conclusion

We propose MAPS, a model-agnostic measure for quantifying the predictability of a multivariable time-series forecasting task given its covariates, lookback window, and forecasting horizon. Across synthetic systems with known predictability, MAPS closely follows the reference predictability, achieving an average linear correlation of $r = 0.963$. Furthermore, across Weather and Influenza-Like Illness forecasting tasks, MAPS shows the strongest relationship with empirical forecasting performance among the evaluated predictability measures, obtaining correlations of $r = 0.961$ and $r = 0.827$, respectively. Overall, MAPS provides a complementary measure for characterising forecasting tasks and interpreting model performance across heterogeneous datasets.

REFERENCES

T. Aksu, G. Woo, J. Liu, X. Liu, C. Liu, S. Savarese, C. Xiong, and D. Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

C. Bandt and B. Pompe. Permutation Entropy: A Natural Complexity Measure for Time Series. *Phys. Rev. Lett.*, 88:174102, 2002.

N. Beck, J. Dovern, and S. Vogl. Mind the naive forecast! a rigorous evaluation of forecasting models for time series with low predictability. *Applied Intelligence*, 55(6):395, 2025.

P. M. Catt. On the Limits of Prediction: Forecastability Profiles and Information Decay in Time Series. *arXiv preprint arXiv:2603.20546*, 2026.

P. Chang, R. Saravanan, T. DelSole, and F. Wang. Predictability of Linear Coupled Systems. Part I: Theoretical Analyses. *Journal of Climate*, 17(7):1474–1486, 2004.

T. DelSole. Predictability and Information Theory. Part I: Measures of Predictability. *Journal of the Atmospheric Sciences*, 61(20):2425–2440, 2004.

L. Devroye, L. Gyorfi, A. Krzyzak, and G. Lugosi. On the Strong Universal Consistency of Nearest Neighbor Regression Function Estimates. *The Annals of Statistics*, 22(3):1371 – 1385, 1994.

C. Dong, D. Faranda, A. Gualandi, V. Lucarini, and G. Mengaldo. Time-lagged recurrence: A data-driven method to estimate the predictability of dynamical systems. *Proceedings of the National Academy of Sciences*, 122(20):e2420252122, 2025.

J. P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57: 617–656, 1985.

M. Finzi, S. Qiu, Y. Jiang, P. Izmailov, J. Z. Kolter, and A. G. Wilson. From Entropy to Epiplexity: Rethinking Information for Computationally Bounded Intelligence. *arXiv preprint arXiv:2601.03220*, 2026.

S. Gao, G. Ver Steeg, and A. Galstyan. Efficient Estimation of Mutual Information for Strongly Dependent Variables. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 277–286, 2015.

J. Gasteiger, J. Groß, and S. Günnemann. Directional Message Passing for Molecular Graphs. In *International Conference on Learning Representations*, 2020.

T. Gneiting, K. Larson, K. Westrick, M. G. Genton, and E. Aldrich. Calibrated Probabilistic Forecasting at the Stateline Wind Energy Center. *Journal of the American Statistical Association*, 101(475):968–979, 2006.

R. Godahewa, C. Bergmeir, G. I. Webb, R. J. Hyndman, and P. Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

G. Goerg. Forecastable Component Analysis. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 64–72, 2013.

N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140:107–113, 1993.

R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

R. H. Inman, H. T. Pedro, and C. F. Coimbra. Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, 39(6):535–576, 2013. ISSN 0360-1285.

I. T. Jolliffe. *Principal Components in Regression Analysis*, pages 129–155. 1986.

G. Kitagawa. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69: 066138, 2004.

L.-W. H. Lehman, R. P. Adams, L. Mayaud, G. B. Moody, A. Malhotra, R. G. Mark, and S. Nemati. A physiological time series dynamics-based approach to patient monitoring and outcome prediction. *IEEE J Biomed Health Inform*, 19(3):1068–1076, 2014.

A. L. M. Levada, F. Nielsen, and M. F. C. Haddad. Adaptive k-nearest neighbor classifier based on the local estimation of the shape operator. *arXiv preprint arXiv:2409.05084*, 2024.

E. N. Lorenz. Deterministic Nonperiodic Flow. *Journal of Atmospheric Sciences*, 20(2):130–141, 1963.

C. Park, J. Z. Huang, and Y. Ding. Domain Decomposition Approach for Fast Gaussian Process Regression of Large Spatial Data Sets. *Journal of Machine Learning Research*, 12(47):1697–1728, 2011.

S. M. Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

G. E. Powell and I. C. Percival. A spectral entropy method for distinguishing regular and irregular motion of Hamiltonian systems. *Journal of Physics A: Mathematical and General*, 12(11):2053, 1979.

L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

P. Ranjan, R. Haynes, and R. Karsten. A Computationally Stable Approach to Gaussian Process Interpolation of Deterministic Computer Simulation Data. *Technometrics*, 53(4):366–378, 2011.

S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey. WeatherBench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

O. Shchur, A. F. Ansari, C. Turkmen, L. Stella, N. Erickson, P. Guerron, M. Bohlke-Schneider, and Y. Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2025.

Y. Shu and V. Lampos. DeformTime: capturing variable dependencies with deformable attention for time series forecasting. *Transactions on Machine Learning Research*, 2025.

Y. Shu and V. Lampos. Sonnet: Spectral Operator Neural Network for Multivariable Time Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

C. A. Sims. Macroeconomics and Reality. *Econometrica*, 48(1):1–48, 1980. ISSN 00129682, 14680262.

J. H. Stock and M. W. Watson. Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002a. ISSN 01621459.

J. H. Stock and M. W. Watson. Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002b.

C. J. Stone. Consistent Nonparametric Regression. *The Annals of Statistics*, 5(4):595–620, 1977.

G. Sugihara, B. T. Grenfell, R. M. May, and H. Tong. Nonlinear forecasting for the classification of natural time series. *Philosophical Transactions of the Royal Society of London, Series A: Physical and Engineering Sciences*, 348(1688):477–495, 1994.

F. Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381, 1981. ISBN 978-3-540-38945-3.

A. Terenin, D. R. Burt, A. Artemev, S. Flaxman, M. van der Wilk, C. E. Rasmussen, and H. Ge. Numerically Stable Sparse Gaussian Processes via Minimum Separation using Cover Trees. *Journal of Machine Learning Research*, 25(26):1–36, 2024.

O. Wang, P. Quan, K. Yang, and M. Srivastava. Spectral Predictability as a Fast Reliability Indicator for Time Series Forecasting Model Selection. *arXiv preprint arXiv:2511.08884*, 2025a.

Y. Wang, H. Wu, Y. Ma, Y. Fang, Z. Zhang, Y. Liu, S. Wang, Z. Ye, Y. Xiang, J. Wang, et al. Accuracy law for the future of deep time series forecasting. *arXiv preprint arXiv:2510.02729*, 2025b.

G. Wen, N. I. Fox, and P. S. Market. A Gaussian mixture method for specific differential phase retrieval at X-band frequency. *Atmospheric Measurement Techniques*, 12:5613–5637, 2019.

C. Williams and C. Rasmussen. Gaussian Processes for Regression. In *Advances in Neural Information Processing Systems*, volume 8, 1995.

M. S. Wulfsohn and A. A. Tsiatis. A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics*, 53(1):330–339, 1997. ISSN 0006341X, 15410420.

S. G. Zadeh, V. Shaj, P. Jahnke, G. Neumann, and T. Breitenbach. Towards Measuring Predictability: To which extent data-driven approaches can extract deterministic relations from data exemplified with time series prediction and classification. *Transactions on Machine Learning Research*, 2025.

A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are Transformers Effective for Time Series Forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11121–11128, 2023.

J. Zhang, X. Wen, Z. Zhang, S. Zheng, J. Li, and J. Bian. ProbTS: Benchmarking Point and Distributional Forecasting across Diverse Prediction Horizons. In *Advances in Neural Information Processing Systems*, volume 37, pages 48045–48082, 2024.

Y. Zhang and J. Yan. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *International Conference on Learning Representations*, 2023.

APPENDIX

A DATASETS AND METRICS

A.1 CHOICE OF MODEL PERFORMANCE METRICS

The proposed MAPS score is designed to quantify the intrinsic predictability of a time series under a rolling-window setting in a model-agnostic way. As described in Section 4.2, for real-world datasets, where reference predictability is difficult to quantify directly, we assess the score by comparing it with the performance of a forecasting model on the same task under the metrics of linear correlation (ρ) and scaled MAE (s). Notably, the SMAPE score of ML models has recently been used to benchmark predictability scores (Wang et al., 2025a). In Figure S1, we compare SMAPE with linear correlation (ρ) and scaled MAE (s) for Sonnet (Shu and Lampos, 2026), where each location is shown in a different colour. Within each location, both ρ and s show a negative relationship with SMAPE as expected. Models with higher correlation tend to have lower prediction errors under the same data distribution. However, this relationship becomes weaker when results are evaluated across locations. This suggests that these metrics capture different aspects of forecasting performance: a predictability score that correlates well with ρ and s need not correlate well with SMAPE. For completeness, we additionally report the relationship between the proposed predictability score and target SMAPE in Appendix C.3.

A.2 GENERATION OF COMPLEX SYNTHETIC DATASETS WITH KNOWN PREDICTABILITY

We construct six synthetic time series with known predictability to evaluate the proposed method under controlled conditions. Figure S2 shows the generated series. For the linear Gaussian state-space model (SSM), smooth Gaussian Process (GP), fractional Gaussian noise, and quasi-periodic process, we characterise temporal dependence through autocovariance functions. Throughout this section, $y_t \in \mathbb{R}$ denotes the observed univariate series at time t, and $\gamma(h) = \mathrm{Cov}(y_t, y_{t+h}) \in \mathbb{R}$ denotes its autocovariance function at lag h, which does not depend on t because the process is stationary.

Linear Gaussian state-space model. We first generate synthetic time series using a 2-dimensional linear Gaussian state-space model. Let $z_t = (z_{t,1}, z_{t,2})^\top \in \mathbb{R}^2$ be the latent state at time t. The state transition is given by

$$z_{t+1} = Az_t + w_t, \qquad y_t = z_{t,1} + v_t\,, \tag{S1}$$

where $w_t \sim \mathcal{N}(0, \sigma_w^2 I_2)$ and $v_t \sim \mathcal{N}(0, \sigma_v^2)$ are independent Gaussian noise. The transition matrix is defined as

$$A = \kappa \begin{pmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{pmatrix}, \qquad \omega = 2\pi f\,, \tag{S2}$$

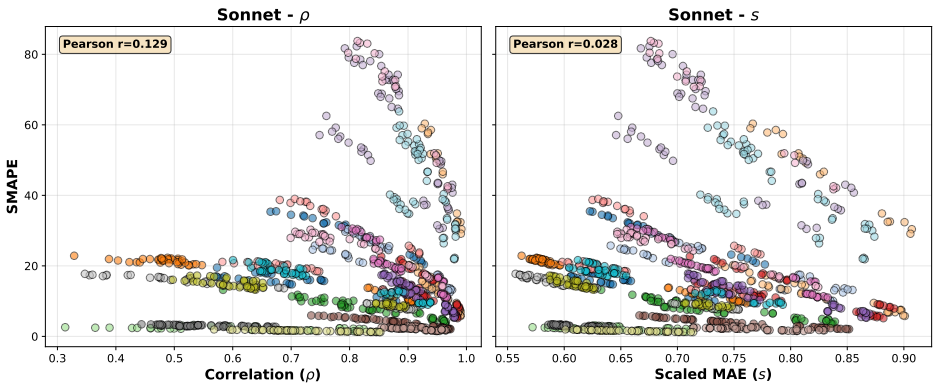


Figure S1: Relationship between SMAPE and linear correlation across weather forecasting tasks for Sonnet against target correlation ρ (left) and scaled MAE score s (right). Each point represents one task for one of 20 locations, with colours indicating different locations.

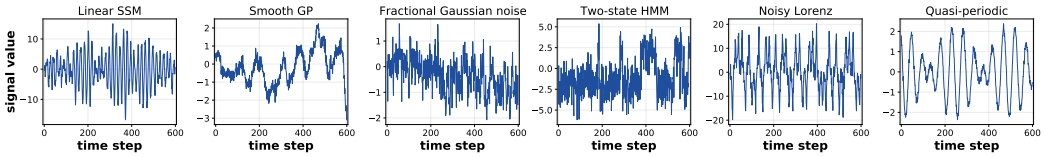


Figure S2: Generated time series for the six synthetic processes with known predictability. Each series contains $T = 4{,}000$ time steps, of which the first 600 are shown.

where f controls the frequency and $\kappa < 1$ controls the rate at which the process loses temporal dependence. This construction produces a stationary, stochastically perturbed oscillatory process with exponentially decaying temporal dependence, while σ_v^2 controls the noise level. In our experiments, we set $f = 0.06$, $\kappa = 0.985$, $\sigma_w^2 = 1$, and $\sigma_v^2 = 0.35$, and initialise with $z_0 = (0,0)^\top$.

Smooth Gaussian Process. The smooth GP is defined by the squared-exponential covariance function, given by:

$$\gamma(h) = \begin{cases} 1 + \sigma_\epsilon^2 & h = 0 \\ \exp\left(-\dfrac{h^2}{2\ell^2}\right) & h \neq 0\,, \end{cases} \tag{S3}$$

where h denotes the temporal lag between two time points. ℓ controls the rate at which temporal correlations decay with increasing time lag, while σ_ϵ^2 controls the variance of the noise, which comes into play only when $h = 0$. In our experiments, we set $\ell = 12$ and $\sigma_\epsilon^2 = 0.05$.

Fractional Gaussian noise process. The next synthetic time series is generated using fractional Gaussian noise, a zero-mean stationary GP with covariance function

$$\gamma(h) = \begin{cases} 1 + \sigma_\epsilon^2 & h = 0 \\ \dfrac{1}{2}\left[(h+1)^{2\eta} - 2h^{2\eta} + |h-1|^{2\eta}\right] & h \neq 0\,, \end{cases} \tag{S4}$$

where h denotes the temporal lag and $\eta \in (0,1)$ controls how much a time series depends on the past values. For $\eta > 1/2$, the process exhibits long-range dependence, with temporal correlations reducing according to a power law. Larger values of η therefore produce stronger and more persistent dependence across long time lags. The term σ_ϵ^2 is the noise variance. We set $\eta = 0.95$ and $\sigma_\epsilon^2 = 0.01$.

Two-state hidden Markov model. The next series is generated by a two-state hidden Markov model with Gaussian emissions. At each time point t, the process is assumed to be in one of two possible states, denoted by $z_t \in \{1,2\}$. The transition probabilities are given by

$$\Pr(z_t = j \mid z_{t-1} = i) = A_{ij}, \qquad A = \begin{pmatrix} \lambda & 1-\lambda \\ 1-\lambda & \lambda \end{pmatrix}, \tag{S5}$$

where $\Pr(\cdot)$ denotes probability, A_{ij} is the probability of transitioning from state i at time $t-1$ to state j at time t, and λ is the probability of remaining in the same state. Given the current state, the time series is generated from a state-specific Gaussian distribution:

$$y_t \mid z_t = i \sim \mathcal{N}(\mu_i, \sigma_i^2)\,, \tag{S6}$$

where μ_i is the mean of state i, and σ_i is its standard deviation. The two hidden states have their own mean and standard deviation, allowing the series to switch between two regimes. We set $\lambda = 0.99$, $(\mu_1, \mu_2) = (-2, 2)$, and $(\sigma_1, \sigma_2) = (1.4, 1.4)$.

Noisy Lorenz system. This series is generated through a stochastic Lorenz state-space model. Let $z_t \in \mathbb{R}^3$ denote the latent state, which evolves according to

$$z_{t+1} = \Phi_{\mathrm{RK4}}(z_t) + w_t, \qquad y_t = z_{t,1} + v_t\,, \tag{S7}$$

where z_{t+1} is the state at the next time point, Φ_{RK4} is the function that takes the current state z_t and computes its next state according to the Lorenz differential equations using the fourth-order Runge–Kutta (RK4) method, $w_t \sim \mathcal{N}(0, \sigma_w^2 I_3)$ is process noise that introduces randomness into the state evolution, $z_{t,1}$ is the first component of the latent state, and $v_t \sim \mathcal{N}(0, \sigma_v^2)$ is the added noise. We use the standard Lorenz parameters $(\sigma, \rho, \beta) = (10, 28, 8/3)$, an integration step of $\Delta t = 0.05$, and noise variance $\sigma_w^2 = 0.02$ and $\sigma_v^2 = 4$. We initialise with $z_0 = (1,1,1)^\top$.

Quasi-periodic process. We next combine several sinusoidal components with frequencies that do not have a common period. This yields

$$y_t = \sum_{j=1}^{J} a_j \cos(2\pi f_j t + \phi_j) + \epsilon_t \,, \tag{S8}$$

where J is the number of sinusoidal components, a_j is the amplitude of component j, f_j is its frequency, ϕ_j is its fixed phase, and $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is independent noise. Because the frequencies do not share a common period, the combined signal does not exactly repeat over time. The underlying signal is fully deterministic because the amplitudes, frequencies, and phases are fixed, while the added noise can make the observed series appear irregular. We set $J = 2$, $(a_1, a_2) = (1, 0.7)$, $f_1 = 0.017$, $f_2 = (\sqrt{5} - 1) f_1$, and $\sigma_\epsilon^2 = 0.01$.

A.2.1 Reference Predictability

Following Chang et al. (2004), we define reference predictability as one minus the variance that remains across possible future sequences given the lookback window, normalised by the total variance in the future target values. For a lookback window of L time steps and a forecasting horizon H, the reference predictability is given by

$$P(L, H) = 1 - \frac{\mathbb{E}\left[\sum_{h=1}^{H} \mathrm{Var}(y_{t+h} \mid y_{t-L+1:t})\right]}{\sum_{h=1}^{H} \mathrm{Var}(y_{t+h})} \,. \tag{S9}$$

For each past input window, we compute how much the possible future target values can vary under the known data-generating process. The variance is computed directly for the 4 GP-based and the hidden Markov models, and approximated with generated futures for the noisy Lorenz system. The reference predictability score is then computed by 1 minus the normalised variance averaged across the sliced windows. $P(L, H) = 1$ thus indicates a completely predictable target, whereas $P(L, H) = 0$ means that observing the past provides no information about the target value. For each synthetic system, we first compute its conditional predictive variance. For the linear Gaussian SSM, smooth GP, fractional Gaussian noise, and quasi-periodic Gaussian process, the conditional variance is available analytically from the autocovariance function $\gamma(h)$ (Williams and Rasmussen, 1995; Park et al., 2011). Let

$$\Sigma_{pq} = \gamma(|p - q|), \qquad c_p = \gamma(H + L - 1 - p) \,, \tag{S10}$$

where $p, q = 0, \ldots, L-1$ index positions in the lookback window. Σ is the covariance matrix of the L past input. $\mathbf{c} = (c_0, \ldots, c_{L-1})^\top$ is a vector that describes how strongly each of the L past input is related to the target target value y_{t+H}. Because the past input and the target are jointly Gaussian, the variance remaining after observing the past is given by:

$$\mathrm{Var}(y_{t+H} \mid y_{t-L+1:t}) = \gamma(0) - \mathbf{c}^\top \Sigma^{-1} \mathbf{c}. \tag{S11}$$

Following Ranjan et al. (2011) and Terenin et al. (2024), we add $10^{-10}\gamma(0)$ to diagonal of Σ for stability.

For each lookback window of L, the forward algorithm (Rabiner, 1989) estimates the posterior probabilities of the two hidden states at the end of the window. The posterior is propagated H steps through the transition matrix, giving future-state probabilities $\pi_{t+H,i}$ in state i at time $t + H$. Using the Gaussian-mixture variance formula of Wen et al. (2019), the predictive variance is given by

$$\mathrm{Var}(y_{t+H} \mid y_{t-L+1:t}) = \sum_{i=1}^{2} \pi_{t+H,i}(\mu_i^2 + \sigma_i^2) - \left(\sum_{i=1}^{2} \pi_{t+H,i} \mu_i\right)^2 . \tag{S12}$$

We average the predictive variance over 400 lookback windows and use the result to calculate the reference predictability score $P(L, H)$.

For the noisy Lorenz system, we generate 500 possible hidden Lorenz states for each lookback window from its target values using a bootstrap particle filter (Gordon et al., 1993; Kitagawa, 1996). We then evolve these states under the known Lorenz dynamics and measure how much their future target values differ. Since the target is the first state variable plus Gaussian noise, the predictive variance is estimated as

$$\mathrm{Var}(y_{t+H} \mid y_{t-L+1:t}) \approx \mathrm{Var}(z_{t+H,1} \mid y_{t-L+1:t}) + \sigma_v^2. \tag{S13}$$

Here, Var is the variance estimated from the propagated particles. We average this variance over 400 lookback windows for obtaining $P(L, H)$.

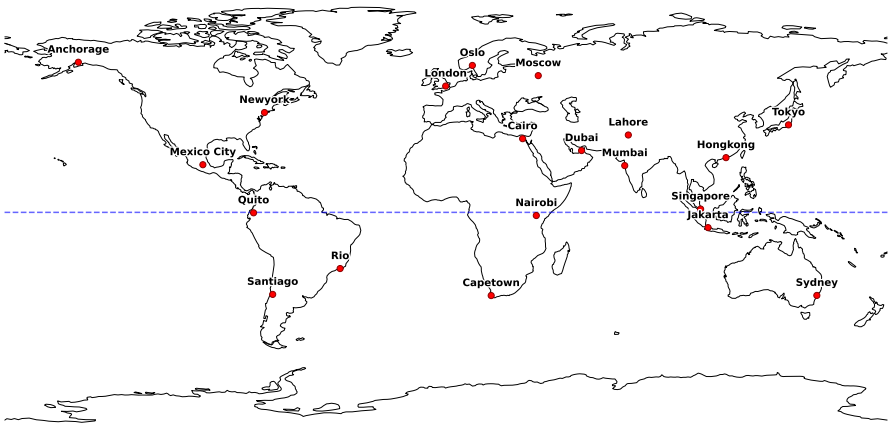


Figure S3: Distribution of the 20 locations used in the weather forecasting tasks.

A.3 Construction of the ILI datasets

Following Shu and Lampos (2025), we construct daily influenza-like illness (ILI) forecasting datasets for England and US Health and Human Services Regions 2, 9, and additionally included region 10. The forecasting target is the ILI rate for each region. ILI rates are obtained from the Royal College of General Practitioners (RCGP) for England and from the Centers for Disease Control and Prevention (CDC) for the US regions. Since the original ILI rates are reported weekly, we linearly interpolate them to a daily resolution. We use daily frequencies of health-related Google search trends as exogenous inputs. For the US regions, search frequencies are collected at the state level and combined to obtain a regional time series. Additionally, we consider the lag in reported ILI rates, specifically 7 days for England and 14 days for the US regions, following the setup in DeformTime (Shu and Lampos, 2025).

A.4 Selection of prediction regions from the WeatherBench dataset

We form the weather forecasting (WEA) dataset using climate information extracted from the WeatherBench repository (Rasp et al., 2020), covering the period from 1979 to 2018, following the data collection process of Sonnet (Shu and Lampos, 2026). Data is sampled from a uniform spatial grid over the globe with a resolution of 5.625° in both latitude and longitude, where each grid point represents a geographical location. To capture a broader range of climatic conditions, we select 20 spatially distant locations distributed across different continents, hemispheres, climate zones, and elevations, as labelled in Figure S3. This allows the dataset to cover diverse weather patterns. For each location, climate data is sampled from the grid point closest to its geographical coordinates. The forecasting target is the 850 hPa temperature in Kelvin (T850), a widely used indicator of atmospheric conditions. Following previous work (Shu and Lampos, 2026), we include 4 additional climate variables as exogenous inputs, namely the 500 hPa geopotential (Z500), 2-metre air temperature (t2m), 10-metre zonal wind (u10), and 10-metre meridional wind (v10). The data is resampled to a temporal resolution of 6 hours. To investigate the contribution of spatial context, we construct two variants for each location. The first contains 45 covariates. It includes all 5 above-mentioned variables from the 8 neighbouring grid points surrounding the target location. This forms a 3×3 spatial window centred on the selected grid point, resulting in 45 input variables. The second contains 5 local climate variables at the target grid point, without accounting for the neighbouring grid points.

B Supplementary Experiment Setup

B.1 Forecasting setup and data splits

Models are trained using Adam with a learning rate of 2×10^{-4}. We set the maximum training epochs to 100 with early stopping based on validation loss, with a patience of 5 epochs. Following

Sonnet (Shu and Lampos, 2026) for WEA, years 2016, 2017, and 2018 are used for testing, with the preceding year used for validation and all earlier data (since 1980) used for training. Following DeformTime (Shu and Lampos, 2025) for ILI, we use 4 test seasons from 2015/16 to 2018/19. The preceding 9 seasons are used for training, and three 60-day periods from the most recent training seasons are selected as a held-out set for validation. All inputs are standardised using training-set statistics. For each task, the predictability score is computed on the training set, while the forecasting performance is evaluated on the test set. All experiments are conducted using a server running Ubuntu 24.04.4, with 2 AMD EPYC 9354 CPUs, 768GB of DDR5 RAM, and 3 NVIDIA L40S GPUs with 48GB VRAM each.

B.2 Details of the ablation components with variance

We describe the variance-based component evaluated in the ablation study in Section 4.6, which replaces the direction-based similarity measure used by MAPS. Specifically, given the neighbourhood $\mathcal{N}_t$ defined in Section 3.1, we compute:

$$v_t = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{k-1}\sum_{j\in\mathcal{N}_t}(y_{j,h}-\bar{y}_{t,h})^2, \qquad \bar{y}_{t,h} = \frac{1}{k}\sum_{j\in\mathcal{N}_t} y_{j,h}. \tag{S14}$$

Here, $y_{j,h}$ denotes the h-th element of the standardised target vector $\mathbf{y}_j$, and $\bar{y}_{t,h}$ is its mean across the neighbours of window t. Thus, v_t measures the variance of the neighbouring targets averaged across the H forecast steps. We define the window-level predictability as

$$u_t = 1 - \min(v_t, 1)\,, \tag{S15}$$

and the overall variance-based score as

$$\mathcal{P}_{\text{var}} = \frac{1}{N}\sum_{t=1}^{N} u_t. \tag{S16}$$

A smaller v_t indicates that neighbouring input windows have more clustered future targets and therefore yielding a higher predictability score.

B.3 Summary of baseline predictability methods

We compare against 6 baseline methods for dataset predictability.

Persistence (Naïve). The persistence baseline predicts target values with the last available observation. We thereby define persistence predictability as the reduction in absolute error relative to predicting the mean target value:

$$P = 1 - \frac{\sum_{t=1}^{N}\sum_{h=1}^{H}|y_{t+h}-y_t|}{\sum_{t=1}^{N}\sum_{h=1}^{H}|y_{t+h}-\bar{y}|}, \qquad \bar{y} = \frac{1}{NH}\sum_{t=1}^{N}\sum_{h=1}^{H} y_{t+h}.$$

Forecastable Component Analysis (ForeCA). Proposed by Goerg (2013), ForeCA defines the predictability of a univariate time series as the complement of the normalised Shannon entropy of its power spectral density. We apply this measure to the complete target series for each training split and report the average across splits.[2]

Spectral predictability. Proposed by Wang et al. (2025a), this method quantifies predictability as one minus the normalised spectral entropy of the power spectrum of a univariate time series. For each training split, we compute the score over the complete series of each covariate and the target separately and average the resulting scores across variables. We then report the average across training splits.

Pattern Complexity (PC). Proposed by Wang et al. (2025b), PC measures the variation in frequency-domain patterns across forecasting windows, with larger values indicating greater complexity. For each input variable, including the target history, we concatenate its L-step lookback window with

[2]Our implementation uses the ForeCA package.

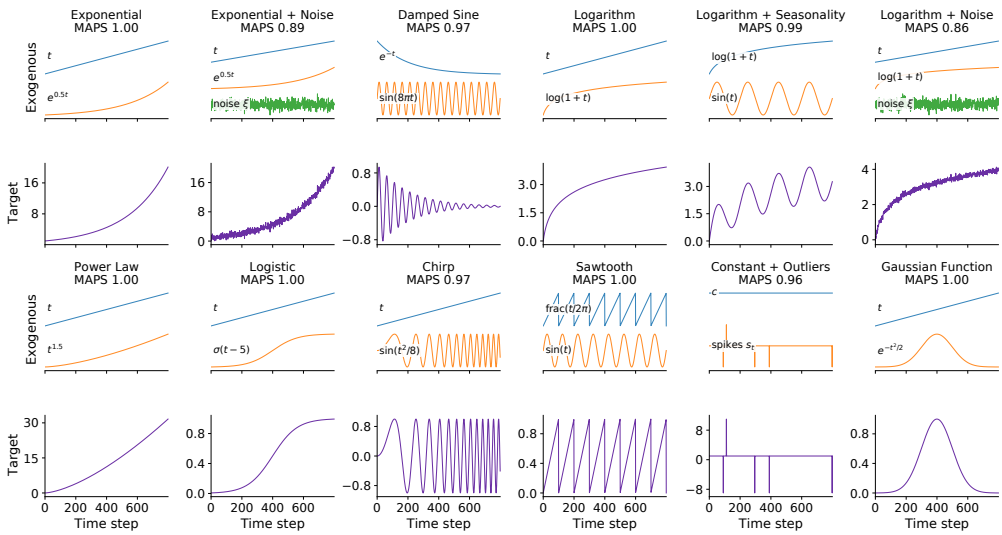


Figure S4: Additional synthetic data and corresponding MAPS scores. Exogenous variables are shown in the first and third rows, with the corresponding target series directly below.

the corresponding H-step future target, compute PC across the resulting windows, and average the scores across variables. We then report the average across training splits.

Mutual Information (MI). Motivated by the predictability measurement proposed in DelSole (2004), we adapt mutual information to multivariable forecasting tasks by constructing paired samples from rolling windows, with an L-step look-back window as input and an H-step target horizon as output. We estimate MI for each input feature using Kraskov et al. (2004) and report the average score.

Epiplexity. Following Finzi et al. (2026), epiplexity measures the learnable information in a dataset, estimated as the area between the training loss curve and the final loss. We approximate epiplexity using ridge regression as a baseline measure of dataset predictability, with higher values indicating more learnable structure. Notably, since epiplexity is model-dependent, it may yield different values under different architectures.

C Supplementary Results

C.1 Results with synthetic data

To validate whether MAPS captures predictability across series with broader patterns, we construct additional synthetic data as shown in Figure S4, with detailed results reported in Table S1. We show rows of exogenous-variable plots, each followed by the corresponding target-variable plots. Each dataset contains one or more exogenous variables, with the target sequence generated as a known function of the inputs. These data include exponential, logarithmic, and periodic patterns such as the chirp, sawtooth, etc. It can be observed that MAPS assigns scores close to 1 across these deterministic systems. In comparison, Spectral produces consistently high scores of approximately 0.92, while ForeCA remains near 0.45. MI correctly identifies the monotonic deterministic functions but assigns substantially lower scores to the damped sine and chirp signals (0.27 and 0.56, respectively). Additionally, we include a constant series with sparse outlier spikes. This series receives a

Table S1: Predictability scores of bounded baseline methods on additional synthetic time series. All scores lie in $[0,1]$, with higher values indicating greater estimated predictability.

Method	Exponential	Exponential + Noise	Damped Sine	Logarithm	Log + Seasonal	Log + Noise	Power Law	Logistic	Chirp	Sawtooth	Constant + Outliers	Gaussian Function
MAPS	0.998	0.888	0.965	0.998	0.990	0.864	0.998	0.999	0.975	0.996	0.965	0.997
ForeCA	0.453	0.100	0.463	0.453	0.455	0.101	0.453	0.453	0.461	0.453	0.789	0.453
Spectral	0.919	0.520	0.918	0.919	0.919	0.521	0.919	0.919	0.918	0.916	0.898	0.919
MI	1.000	0.580	0.268	1.000	0.756	0.562	1.000	1.000	0.556	1.000	0.003	1.000

Table S2: Subtraction-corrected MAPS predictability score analysis w.r.t. different neighbourhood size k, averaged over 20 locations.

k	L	28			56			84		
	H	4	12	28	4	12	28	4	12	28
1		.031	.045	.052	.026	.036	.045	.025	.036	.044
2		.283	.255	.237	.264	.243	.229	.259	.238	.228
8		.433	.396	.367	.411	.381	.359	.401	.373	.356
12		.464	.424	.393	.442	.409	.384	.432	.401	.382
32		.513	.468	.434	.492	.454	.425	.484	.447	.423
48		.525	.479	.443	.504	.465	.435	.496	.459	.433
64		**.530**	**.483**	**.448**	.510	.471	.440	.503	.465	.438
256		.529	.482	.447	**.514**	**.474**	**.442**	**.508**	**.469**	**.441**
1024		.480	.440	.408	.473	.437	.408	.468	.432	.404
2048		.416	.382	.354	.414	.383	.357	.411	.379	.355
4096		.185	.169	.155	.184	.170	.158	.184	.169	.157
all		.000	.000	.000	.000	.000	.000	.000	.000	.000

Table S3: Linear correlation between MAPS score and Sonnet's forecasting performance, taken across the 20 regions within each setting.

k	L	28			56			84		
	H	4	12	28	4	12	28	4	12	28
1		**.942**	.943	.937	**.932**	.958	.950	**.943**	.954	.948
2		.879	.971	.958	.849	.976	.966	.836	.976	.967
8		.856	.976	.967	.820	.977	.972	.804	.976	.973
12		.852	.976	.968	.815	.977	.972	.798	.976	.973
32		.846	.977	.970	.810	.977	.973	.793	.976	.975
48		.844	.977	.971	.808	.977	.974	.794	.976	.976
64		.843	**.978**	.971	.809	.978	.975	.794	.976	.976
256		.845	.976	**.972**	.820	**.980**	**.979**	.807	**.977**	**.980**
1024		.840	.954	.956	.827	.967	.969	.824	.961	.972
2048		.803	.903	.912	.799	.921	.928	.803	.913	.935
4096		.764	.837	.853	.748	.853	.865	.747	.837	.872
all		.360	.424	.433	.342	.428	.441	.339	.430	.415

high MAPS score of 0.97. This result indicates that infrequent irregularities have a limited effect on the overall predictability of the series. Since the underlying series remains largely constant, the outliers introduce limited uncertainty relative to the overall signal.

C.2 Further analysis on the number of nearest neighbours (k) in MAPS

We use the `WEA` dataset to assess how the neighbourhood size k affects the predictability score. Results are enumerated in Table S2. For each of the 20 weather regions and 9 forecasting tasks, we evaluate the effect for values of k ranging from 1 to all available neighbours and average the results across regions. To compare results across different values of k, we remove the similarity that occurs by chance. Specifically, we randomly shuffle the target vectors $\mathbf{y}_t$ across the input samples while keeping the input vectors $\mathbf{z}_t$ and their neighbourhoods $\mathcal{N}_t$ unchanged. We then recompute the predictability score using the shuffled targets and report $\mathcal{P}_{\text{subtraction}} = \mathcal{P}(\mathbf{X} \to \mathbf{y}) - \mathcal{P}(\mathbf{X} \to \mathbf{y}_{\text{shuffled}})$. The shuffled score measures the neighbourhood similarity expected when the inputs and their futures are unrelated. Subtracting it removes the chance similarity associated with each value of k.

This subtraction is particularly important for small values of k. When $k = 1$, each input window is evaluated only on the target vector of its nearest input neighbour. Increasing k provides a more stable estimate of local future behaviour by combining similarity scores from multiple nearby inputs. At $k = 8$, the original MAPS score exceeds the permutation baseline by at least 0.35 across all forecasting tasks, showing that it captures substantially more structure than $\mathcal{P}_{\text{subtraction}}$. Overall, intermediate values of k provide the clearest predictability signal: overly small neighbourhoods offer limited evidence beyond chance, whereas very large neighbourhoods dilute the local relationship between inputs and their futures.

We also examine how well MAPS reflects forecasting performance across the same range of k in Table S3. Specifically, we compute the linear correlation r between MAPS and Sonnet's forecasting performance ρ. The correlation remains high across a broad range of k, including our default choice $k = \lfloor \log_2(N) \rfloor = 12$. Interestingly, the three $H = 4$ tasks reach the highest correlation at $k = 1$, likely because the target prediction of each input sample is compared only with that of its closest neighbour, preserving small differences in short-term dynamics. On the other hand, short-horizon targets are also easier to predict, reflected in the forecasting performance. Meanwhile, most longer-horizon settings peak at moderate values of k. This indicates that the default setting of k provides a robust choice without requiring k to be tuned separately for each task.

C.3 Results with additional metrics

Figure S5 compares the predictability scores with Sonnet's SMAPE. For the reasons discussed in Section A.1, we report these results for completeness rather than use SMAPE as the primary reference. It can be observed that all methods correlate poorly with SMAPE, and stronger alignment with SMAPE does not necessarily indicate a more accurate predictability measure.

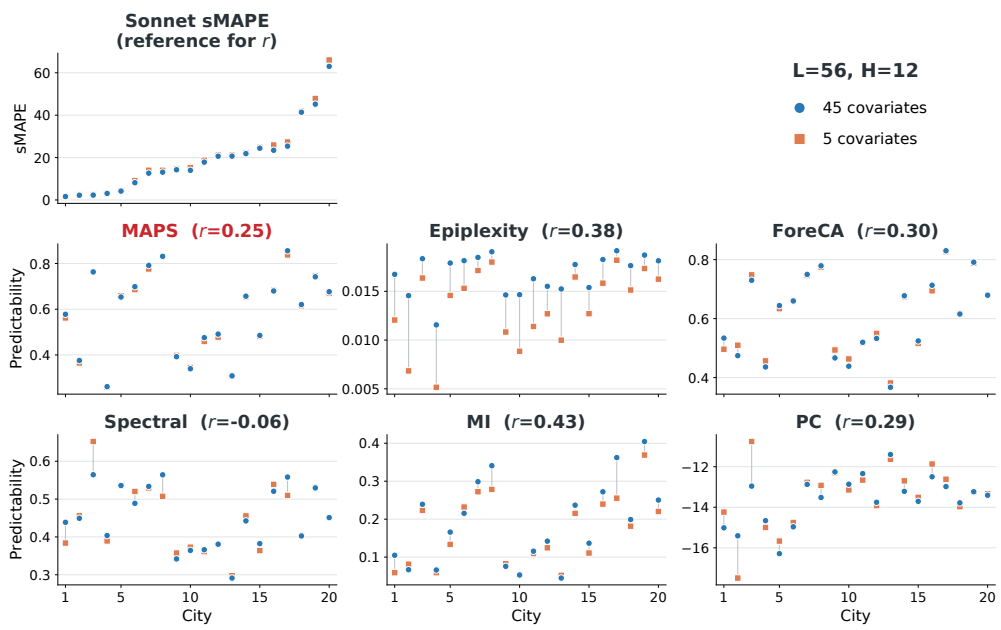


Figure S5: Relationship between different predictability measures and the target SMAPE across all evaluated locations under the Sonnet forecasting model.

C.4 Results with additional machine learning model

To verify that the analysis of MAPS is not limited to a single forecasting model, we further provide results with Crossformer Zhang and Yan (2023) in place of Sonnet. Table S4 reports the linear correlation r across 20 locations between each predictability score and Crossformer's performance, measured by ρ and s. When comparing the predictability scores with Crossformer's target correlation ρ, MAPS yields the strongest correlation r in 8 of the 18 configurations. It performs best in five of the six configurations with $H = 12$ and three of the six configurations with $H = 28$. Epiplexity is the most competitive baseline, leading all six configurations with $H = 4$ and three configurations with $H = 28$, while ForeCA leads the remaining $H = 12$ configuration. Turning to the scaled MAE score s, MAPS yields the strongest correlation in 13 of the 18 configurations, including five of the six configurations at both $H = 12$ and $H = 28$. ForeCA leads the remaining five configurations. These results show that the strong alignment between MAPS and forecasting performance is not limited to Sonnet.

Table S4: Linear correlation r between predictability scores and Crossformer's forecasting performance, measured by target correlation ρ or scaled MAE s, across the 20 WEA locations. Predictability scores and model performance are averaged over the 2016–2018 test periods before computing r. Results are reported across lookback windows (L), forecasting horizons (H), and covariate numbers (C). The best results are bolded.

	Method	$L=28$						$L=56$						$L=84$					
		$H=4$		$H=12$		$H=28$		$H=4$		$H=12$		$H=28$		$H=4$		$H=12$		$H=28$	
		$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$
ρ	MAPS	.847	.898	**.973**	.966	**.970**	.965	.811	.858	**.977**	**.972**	**.971**	.970	.800	.847	**.974**	**.975**	**.970**	.969
	Epiplexity	**.989**	**.959**	.957	.947	.968	**.970**	**.988**	**.953**	.926	.945	.943	**.971**	**.958**	**.956**	.860	.951	.858	**.971**
	ForeCA	.805	.838	.972	**.969**	.967	.968	.803	.829	.975	.970	.969	.967	.810	.841	.972	.973	.964	.966
	Spectral	.467	.513	.831	.764	.850	.774	.463	.506	.837	.775	.853	.785	.477	.516	.843	.774	.869	.786
	MI	.746	.782	.901	.864	.910	.880	.718	.750	.904	.870	.922	.884	.718	.750	.903	.877	.922	.883
	PC	.348	.461	.041	.311	-.017	.301	.311	.452	-.065	.222	-.070	.219	.297	.453	-.118	.206	-.142	.193
s	MAPS	**.924**	**.944**	**.961**	.946	**.955**	.949	.913	**.934**	**.965**	**.952**	**.955**	**.952**	.905	.929	**.968**	**.959**	**.950**	**.956**
	Epiplexity	.918	.904	.911	.893	.938	.935	.913	.902	.882	.886	.915	.936	.876	.903	.832	.891	.836	.937
	ForeCA	.916	.932	.958	**.954**	.950	**.950**	**.920**	.931	.958	.951	.950	.949	**.920**	**.935**	.957	.955	.940	.951
	Spectral	.680	.665	.904	.811	.898	.825	.681	.666	.905	.818	.898	.826	.692	.668	.910	.821	.904	.827
	MI	.898	.885	.911	.856	.909	.871	.888	.876	.920	.865	.924	.879	.885	.871	.920	.877	.919	.883
	PC	.318	.413	.042	.281	-.007	.253	.215	.349	-.094	.175	-.079	.176	.161	.315	-.156	.140	-.145	.149

Table S5: Linear correlation r between target-only predictability scores and Sonnet's forecasting performance. Arrows indicate whether r increases or decreases relative to the corresponding result in Table 1 and = indicates no change.

	Method	$L=28$						$L=56$						$L=84$					
		$H=4$		$H=12$		$H=28$		$H=4$		$H=12$		$H=28$		$H=4$		$H=12$		$H=28$	
		$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$	$C=45$	$C=5$
ρ	MAPS	.822↓	.854↓	.974↓	.964↓	.960↓	.954↓	.787↓	.827↓	.974↓	.963↓	.965↓	.952↓	.764↓	.812↓	.974↓	.965↓	.967↓	.962↓
	ForeCA	.801↑	.836↑	.972↑	.974↑	.961↓	.972↑	.804↑	.845↑	.976↑	.974↑	.970↑	.963↑	.805↑	.850↑	.974↑	.972↑	.973↓	.973↑
	Spectral	.806↑	.840↑	.968↑	.970↑	.954↑	.967↑	.808↑	.850↑	.973↑	.970↑	.963↑	.958↑	.810↑	.854↑	.970↑	.969↑	.968↑	.969↑
	MI	.713↓	.759↓	.940↑	.940↑	.944↑	.953↑	.703↓	.757↓	.934↑	.936↑	.948↑	.948↑	.701↓	.755↓	.931↑	.935↑	.951↑	.956↑
	PC	.154↓	.129↓	-.097↓	-.073↓	-.166↓	-.115↓	.015↓	-.014↓	-.229↓	-.212↓	-.263↓	-.232↓	-.031↓	-.065↓	-.318↓	-.301↓	-.287↓	-.278↓
s	MAPS	.906↓	.933↓	.968↑	.964↑	.960↑	.955↑	.894↓	.919↓	.973↑	.962↑	.957↓	.955↓	.878↓	.911↓	.967↑	.958=	.956↓	.954↓
	ForeCA	.915↑	.936↑	.956↓	.960↓	.947↓	.952↓	.922↑	.938↑	.965↓	.959↓	.953↓	.955↓	.921↑	.940↑	.958↓	.956↓	.953↓	.954↓
	Spectral	.919↑	.940↑	.954↑	.957↑	.943↑	.950↑	.926↑	.941↑	.964↑	.956↑	.950↑	.953↑	.924↑	.944↑	.956↑	.953↑	.950↑	.953↑
	MI	.888↓	.934↑	.987↑	.988↑	.979↑	.982↑	.889↓	.929↑	.987↑	.986↑	.983↑	.984↑	.886=	.924↑	.984↑	.981↑	.986↑	.986↑
	PC	.176↓	.131↓	-.060↓	-.046↓	-.156↓	-.098↓	-.010↓	-.060↓	-.196↓	-.203↓	-.239↓	-.217↓	-.085↓	-.136↓	-.299↓	-.291↓	-.274↓	-.252↓

C.5 Additional analysis on predictability with univariate setup

We further evaluate all methods in a univariate setting, where only the target series is used as input. The resulting scores are compared with Sonnet's performance using $C=45$ and $C=5$ inputs. This analysis therefore evaluates how much information about forecasting difficulty is lost when the predictability measures exclude the external covariates. The results are presented in Table S5. Compared with the multivariable results in Table 1, the correlation between MAPS and ρ decreases in all 18 configurations. This suggests that the exogenous covariates provide useful information about forecasting difficulty. Spectral improves in every configuration, MI generally improves at longer horizons, PC performs worse, and ForeCA changes only slightly.

C.6 Runtime Comparison

Figure S6 compares the median per-window runtime. We change one factor at a time over $C \in \{4,16,64,256,1024\}$, $L \in \{7,14,28,56,112,224\}$, and $H \in \{1,4,7,14,28,56,112\}$, holding the others at $L=28$, $H=7$, and $C=64$. Across the evaluated settings, the runtime of MAPS does not change significantly because its dominant nearest-neighbour search operates in the fixed d-dimensional PCA space. By contrast, all baselines become more computationally expensive as the number of variables increases. Additionally, MI also requires more runtime for longer lookback windows.

D Limitations

The main limitation is that MAPS estimates predictability relative to the observed data distribution and should not be interpreted as a theoretical upper bound on forecasting performance. As demonstrated under distribution shift, a score computed on training data may not characterise an unseen deployment distribution. Furthermore, MAPS evaluates directional consistency and does not capture differences in forecast magnitude. Finally, although we evaluate synthetic systems and real-world weather and ILI tasks, further experiments across a broader range of domains are required to establish its general applicability.

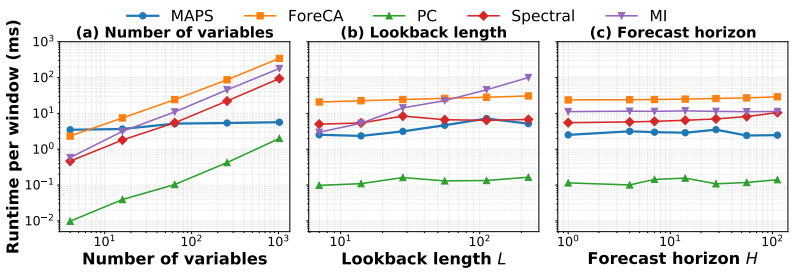


Figure S6: Runtime analysis of predictability estimators. We measure the median per-window runtime (ms) versus the number of variables, lookback window length, and forecast horizon.